Plant Archives Volume 20 No. 2, 2020 pp. 4245-4250 **e-ISSN:2581-6063 (online), ISSN:0972-5210**

HADOOP BIG DATA MINING: AN EFFECTIVE MAPREDUCE TOOL FOR CLASSIFYING SUGARCANE YIELD DATA

R. Revathy[1], P. Murali[2*] and S. Balamurali[1]

[1]Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil-626126 (Tamil Nadu) India.
[2*]Economics and Statistics Section, ICAR-Sugarcane Breeding Institute (SBI)–GOI, Coimbatore-641007 (Tamil Nadu) India.

Abstract

India being named as the kingdom of agriculture, its financial system is predominantly based upon the agricultural crop yield and its products. Farmers previously predict the yield of certain crops through their own practice and observed weather parameters. Data Mining is one of the originating research fields in analyzing the prediction of crop yield since it is an effective method to discover unseen information from the influence of the existing climatic dataset. The purpose of this work is classifying the sugarcane yield data by means of implementing various decision tree algorithms where the classes are categorized as low, moderate and high. The decision algorithms like C4.5, C5.0 and Random forest are carried out through the Hadoop framework by the way of MapReduce implementation and compared the algorithms via evaluating accuracy and error rate. MapReduce based decision tree techniques prove the novelty of this present study on account of classifying the yield data competently. The findings of this research may out the farmers for predicting the sugarcane crop yield in future for market mobility.

Key words: Hadoop, BigData mining, MapReduce decision tree, Climatic data, Sugarcane yield.

Introduction

Nowadays the data stored in the database are frequently increasing every day (Revathy *et al*., 2019). BigData is an exact huge volume of data that is created from various sources like either structured or unstructured or semi-structured format. BigData mining simply transforms the unprocessed data into valuable information. Therefore, it acts as a significant methodology in all emerging field of study with rising value. The most important components of BigData are termed as Volume, Velocity, Variety and Veracity (4V's). Volume represents the data size where velocity defines the processing speed of the data. Variety describes the different types of data and veracity identifies the accuracy prediction that is verified by testing the algorithm (Sahu *et al*., 2019).

The data produced from the agricultural sectors are remote sensing-based, satellite-based and practice of farming manually. Farmers are reaping not only the agricultural crops but also producing vast data for future analysis. The yield of the crops is entirely depending upon the conditions of the weather, precipitation, pests and diseases (Veenadhari *et al*., 2011). Since the profitable system of India is partially belonging to the Agricultural field, data mining in agriculture plays an important role in predicting the yield. In this research paper, BigData mining algorithms has been explored the climatic data in order to classify the sugarcane yield data into three classes like low, moderate and high.

Best crop production method is a way of precision agriculture that is getting better information about crops using machine learning algorithms (Rale *et al*., 2019). According to (Aksu and Dogan, 2019), data mining algorithms like Naïve Bayes, J48, Random forest, artificial neural network and decision trees were implemented using weather and crop parameters. After applying prediction rules, crop information like nature of the crop, pests and yield were classified and predicted.

The prediction of the crop yield is constantly being a challenging task. A model-based system was designed to

***Author for correspondence :** E-mail: p.murali@icar.gov.in

classify the crop production by means of observed data. Even though many machine learning algorithms and other techniques are applied, random forest method obtained good results (Rale *et al*., 2019). Integrating the output of various models with ensembling booster could improve the entire model performance and eliminate the overfitting of the decision tree.

Climatic variables have become an essential influence on the growth of sugarcane. Random forest and gradient boosting algorithms served better forecasting which directed a way to improve the decisions. The models integrated with climatic forecast could achieve in finding the seasonal climatic variations earlier and the farmers could regulate the expectation in sugarcane production (Revathy and Lawrance, 2017).

Relief feature selector algorithm was worked on crop pest data for selecting choosy features in place of feeding full features to the model. The algorithm could appropriately find the weight of the attributes with strong dependency and help in eliminating the redundant attributes (Rosario and Thangadurai, 2015). Identifying significant features in the dataset could certainly improve the classification accuracy. In order to execute the agricultural crop yield data in (Krishna Kumar and Kiruthika, 2015), ID3, C4.5 and C5.0 decision tree algorithms were implemented and compared. C5.0 performed effectively in big data set which could handle all type of data like continuous, categorical, date, etc and also eliminate missing values.

The rice diseases were classified with its symptoms using a support vector machine (SVM) with Radial Basis Function (RBF) method. It was revealed that SVM based RBF had provided an excellent classification for predicting rice diseases (Revathi *et al*., 2011). The researchers have carried out C4.5 decision tree using Hadoop MapReduce system for compacting the size of the data. While processing the MapReduce based decision tree algorithm, the rules for the dataset were generated by performing mapper and reducer methods.

Mapper and reducer are the two functions used for parallel processing of data instead of sequential practice (Yang and HiongNgu, 2016). Map function feed the data and class into the attribute table represented as keys and values. Reduce function further merge the data with related class values and then keep those data at count table. C4.5 and maximum similarity tree (MST) based MapReduce algorithms were discussed by (Glory *et al*., 2015). The entropy is calculated in C4.5 for attribute selection but in MST the attributes are selected by calculating similarity of the attributes. That's why MST granted excellent performance as compared with C4.5 decision tree.

The research focuses on assessing the significance of climatic parameters on the sugarcane yield. Decision trees implementing MapReduce has revealed that there must be a correlation between climatic parameters and sugarcane productivity. The proposed research work is classifying 23 years (1996-2018) of sugarcane yield data with the influence of weather variables by using Hadoop MapReduce based C4.5, C5.0 and Random forest decision tree classification algorithms. The rules generated by the algorithms proved that the most influenced parameter in this dataset is rainfall and consecutively predicted the classes responsible for low, moderate, high sugarcane crop productivity under certain climatic circumstances. Hence, the study was conducted with the following objectives

1. Feature selection
2. MapReduce Classification
3. Accuracy Prediction

Materials and Methods

Relief feature selection

Feature selection is an essential method in data mining algorithms to enhance the quality of the model in terms of accuracy. It discovers the feature subset by eliminating unwanted or redundant attributes. The significance of selecting features minimizes the cost of learning of the algorithm and improves its performance (Revathy *et al*., 2019). This research includes relief feature selection to improve the interpretability of the classification and reduces the training time by avoiding overfitting of the decision trees.

Relief feature selection was first developed by Kira and Rendell (Rosario and Thangadurai, 2015), which can determine the weights of both continuous and discrete attributes based on distance between instances. The average weight of each attribute is obtained by relief algorithm. It was recognized as an easy and efficient feature selection for weighing features as compared with other feature selection methods. The output ranges of the Relief algorithm are stated between -1 and 1 for each feature (Sutha and Tamilselvi, 2015).

Relief feature selection pseudocode

Input: Set of attribute values and the class values; Output: the vector We of estimations of the qualities of attributes; set Wt[a] = 0 for each attribute a; for i = 1 to n do; select sample sa from data at random; find nearest hit sa_h and nearest miss sa_m; for j = 1 to a do; $\Delta Wt_i[a]$ =

diff(a, sa_i, sa_m) - diff(a,sa_j, sa_h); Wt[a] = Wt[a] +ΔWt_i[a]; end for; end for; for each attribute a; do;Wt[a] = Wt[a]/ n; end for; where diff(a, sa_i, sa_j) = 0, if sa_i[a] = sa_j[a]; = 1, if sa_i[a] ≠ sa_j[a]

From the set of seven features, the feature that holds the highest weight is selected as a splitting attribute for tree generation. Since the rainfall attribute enfolds the maximum gain information, it has been preferred as a splitting attribute for MapReduce based decision tree construction. Depending upon the rainfall data value, the classes of sugarcane crop yield are classified using C4.5, C5.0 and Random forest classifiers.

MapReduce based C4.5, C5.0 and Random forest decision tree

This research proposed MapReduce implementation of C4.5, C5.0 and Random forest decision tree in the form of Map and Reduce methods. Before implementing decision tree algorithms, the first phase of relief feature selection was executed to filter the unimportant attributes (Bikku *et al*., 2016). The filtered attributes are sending over to the mapper function and the corresponding dataset is processed by creating many small chunks of data. The second phase implemented classification algorithms in

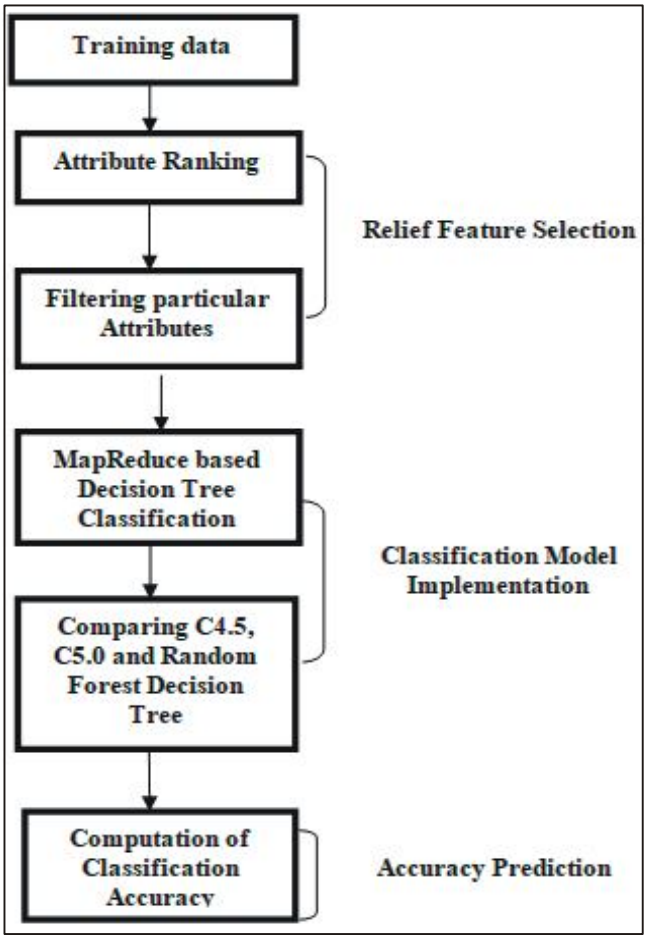

Fig. 1: Three phases of the research workflow.

Table 1: Filtering of attributes by Relief filter.

No.	Attribute Name	Weight
1	Temperature	0.20101457
2	Relative Humidity	0.10450000
3	Wind Speed	0.14573257
4	Pressure	0.16006307
5	Short-wave irradiation	0.06436666
6	Solar Radiation	0.08654994
7	Rainfall	0.95433387

the type of tree formation based on the framework of Hadoop MapReduce. Mapper function identifies the presence of attributes and processed keys and values as pair format. After recognizing the attributes set, the mapper function computes information by splitting of attributes. Reduce function finds the important attributes and then build the decision tree by the arrangement of shufûing and merging procedures. The following fig. 1 describes the three phases of the entire research implementation of this study (Revathy and Lawrance, 2017).

C4.5 classifier can hold continuous as well as discrete features. In order to execute continuous features, the C4.5 classifier sets a boundary value and then divides the data according to its values. Despite C4.5 and C5.0 is able to generate classification either in the form of decision trees or rule sets, C5.0 is efficient in terms of processing speed, memory and accuracy. C5.0 algorithm extended the version of C4.5 algorithm which was designed by Ross Quinlan in 1993 (Dai, 2014). MapReduce based C5.0 classifier constructs the tree by splitting the attributes recursively. Unlike C4.5, C5.0 eliminates the missing data at the time of tree construction and it maintains boosting by prune the unwanted branches of the trees. Although C5.0 is flexible for handling big data sets, the speed and size are substantial. Therefore,

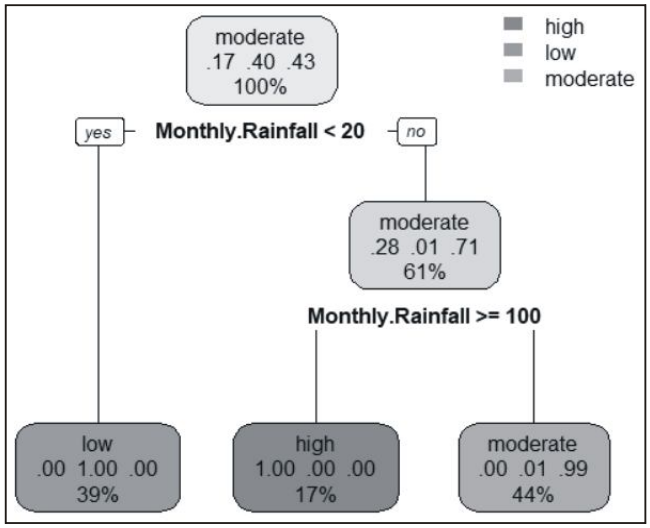

Fig. 2: Construction c MapReduce based C4.5 Algorithm.

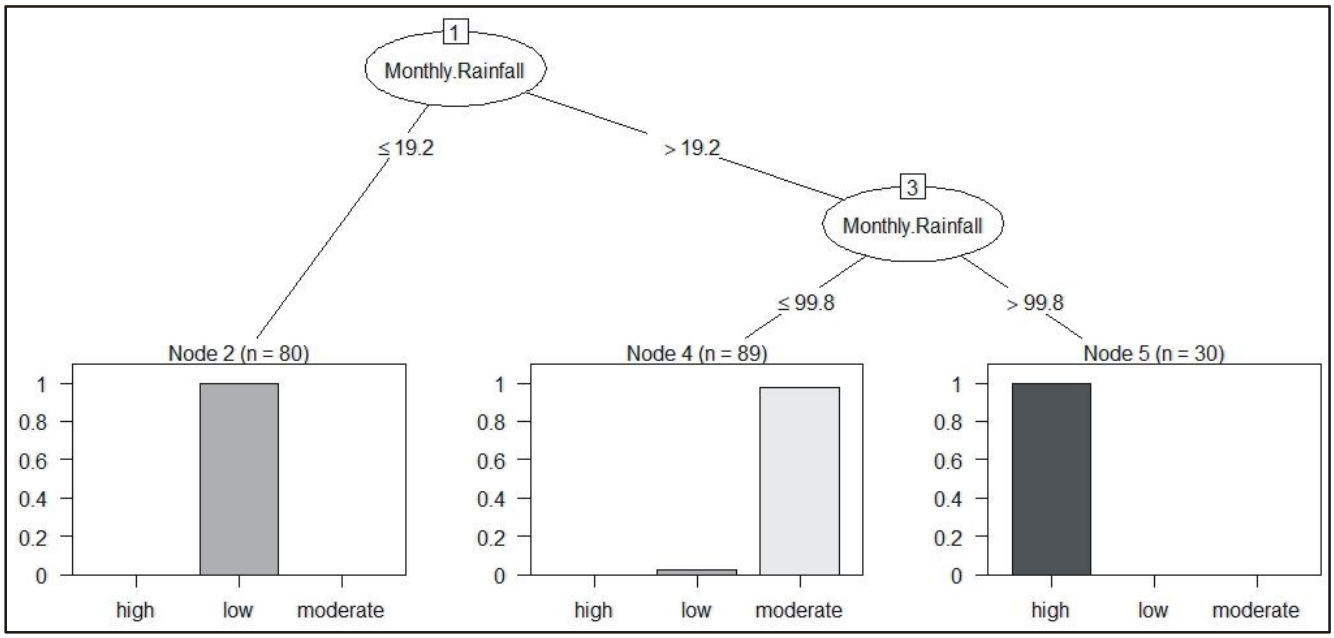

Fig. 3: Construction of MapReduce based C5.0Algorithm.

MapReduce based C5.0 decision tree minimizes the time and space complexity.

The random forest is a classification method in where thousands of trees are constructed. The training instances are executed for each tree with replacement. The main two concepts of random forest are randomly sampling the training data while constructing trees and randomly attribute subsets are considered while splitting of nodes. All three algorithms are belonging to classification techniques that are supervised in complexion (Yang and HiongNgu; 2016). Even C4.5 and C5.0 are termed as good classifiers, perhaps we still require to protect against overfitting of tree. MapReduce based random forest overcome the overfitting issues by merging or averaging the different trees. It works perfectly with large datasets and possess more accuracy values.

Accuracy Prediction

Accuracy is one of the testing methods to validate the trained classification algorithm. The formal accuracy prediction is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}} \quad (1)$$

Sensitivity

Sensitivity is a key testing metric to test the classification algorithm. It is otherwise termed as recall and can be calculated by the following formula:

$$Sensitivity = \frac{No.\ of\ true\ positives}{No.\ of\ ture\ positives + No.\ of\ false\ negatives} \quad (2)$$

Specificity

Specificity is also called as a true negative value. Since the equation does not have false negative and true positive, specificity may acquire a biased result, particularly for imbalanced classes. It is calculated as:

$$Specificity = \frac{No.\ of\ negatives}{No.\ of\ ture\ negatives + No.\ of\ false\ positives} \quad (3)$$

Precision

Precision is also called a positive predictive value. As the equation does not have false negative and true negative, precision obtains a biased result, particularly for imbalanced classes. It is calculated as:

$$Precision = \frac{No.\ of\ ture\ positives}{No.\ of\ ture\ positives + No.\ of\ false\ positives} \quad (4)$$

Error Rate

The performance of the classifiers can also express in terms of error measure which is computed by following

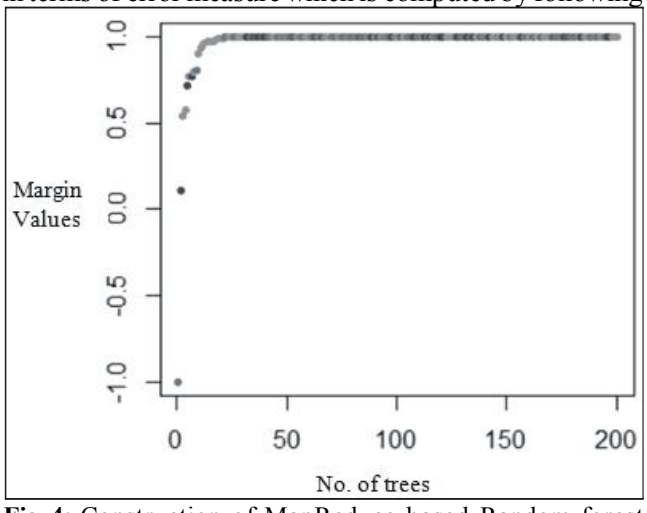

Fig. 4: Construction of MapReduce based Random forest Algorithm.

Table 2: Testing of MapReduce based C4.5, C5.0 and Random forest classification model.

MapReduce Based Decision Trees	Accuracy (in %)	Error (in %)	Precision (in %)	Sensitivity (in %)	Specificity (in %)
C4.5	93.85	4.68	88	97	92
C5.0	96.92	3.08	96	98	97
Random Forest	98.44	1.56	99	99	99

the equation:

$$\text{Error rate} = \frac{\text{Number of misclassification instances}}{\text{Total number of instances}} \quad (5)$$

Results and Discussion

Data interpretation

The study was performed at the Coimbatore district of Tamil Nadu. The data related to this study were collected from the ICAR- Sugarcane Breeding Institute (SBI), Coimbatore. Daily climatic data and annual yield of sugarcane data from the period 1996 to 2018 were accessed for implementing decision tree classification algorithms. The three above mentioned decision trees with MapReduce have been accomplished in Hadoop framework using R programming tool.

Implementation of MapReduce based C4.5, C5.0 and Random forest classification model

Before classifying the sugarcane yield data, this research required relief filter feature selection to select the remarkable features so as to improve the classification accuracy. Relief filter evaluates the weight with its distance among the instances and filtered the features according to its weight obtained in table 1. Since the attribute Rainfall has the maximum weight value; it is elected as root node of the decision tree. The filtered features are then ready to train the decision tree algorithms with MapReduce implementation for classifying and predicting the crop yield. The experimentation of yield data is applied in the R Software tool; data are then processed through MapReduce which is achievable by rmr2 packages.

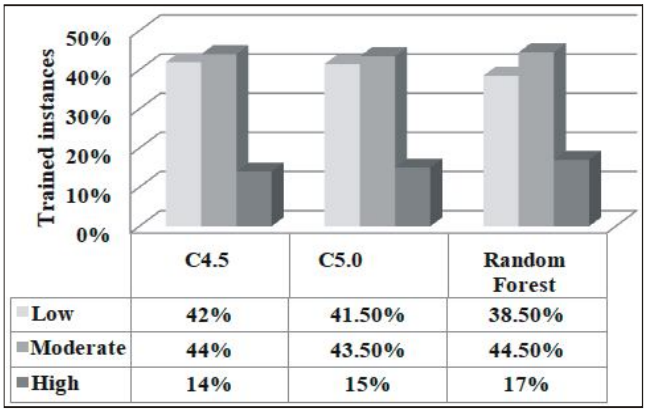

Fig. 5: Level of yield instances classification.

MapReduce based C4.5, C5.0 and Random forest decision trees are built and classified the yield classes as low, moderate and high accordingly for the year 1996-2019. Each algorithm worked proficiently in accordance with its individual abilities. The dataset is initially divided into training data and testing data. The training data helped to construct the decision trees whereas testing data drew the conclusion of accuracy and error rate of the trained algorithm. Mapper and Reduce functions in decision tree algorithms permit to work efficiently on BigData. Fig. 2-4 depicts the three different MapReduce based decision trees in the effective framework of Hadoop for 23 years of climatic and sugarcane yield data.

The massive weather dataset was conceded to the mapper function subsequently processed it and many small chunks of data are created. The decision trees are produced with its rules using the generation of intermediate key and value pair. The following fig. 5 represents how far the classes of the sugarcane yield are classified by low, moderate and high. The sugarcane yield classes were classified and predicted by approximately 40% either low or moderate and the obtained decision tree has found to be three levels of structure.

The C5.0 decision tree in fig. 3 was constructed with its information gain values and eliminated missing values while the branches were segregated, but C4.5 executed with its missing values. Hence C5.0 proved its superiority in fig. 5 as compared with C4.5. Random forest randomly selected the climatic data and important features were identified to build each decision tree. The random subset of the features is only considered here for every node in each tree.

Technically random forests haven't overfitted since we could not draw a conclusion with many trees. Even though the C5.0 decision tree is constructed faster and hold very less memory due to parallel processing, random forest classified the training data with less error. As random forest created many trees, there might not be

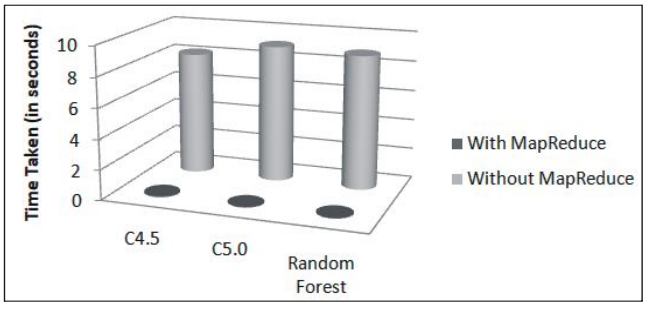

Fig. 6: Construction time of decision trees.

obtaining much error. For random forest algorithm, 200 trees were totally constructed and each tree holds attribute subset. The positive margin values in fig. 3 revealed that the trees effectively build with climatic data and efficiently classified the yield instances.

With the purpose of comparing the testing values obtained from MapReduce based decision trees algorithms, table 2 covers the accuracy rate, error rate, precision, sensitivity and specificity of the classification. From the derived results MapReduce based random forest offered high accuracy, precision, sensitivity and specificity with less error rate as compared with MapReduce based C4.5 and C5.0 decision tree algorithms.

Fig. 6 displayed the construction time of decision trees with and without the implementation of MapReduce. Because of parallel processing, MapReduce based decision trees improved its time and memory management while judging against without MapReduce algorithms. It was proved that the MapReduce decision trees classified the yield classes with minimum duration than normal decision tree algorithms.

Conclusion

The study proposed MapReduce based different classification algorithms for classifying the sugarcane yield classes from accessible observed climatic parameters. It was clearly identified from the results that there was a correlation between climatic circumstances and sugarcane productivity. The most influential of the rainfall parameter on the sugarcane yield was confirmed by rule accuracy of the classification. Decision trees with the implementation of MapReduce were more precise and faster to execute the algorithm in predicting the yield classes. As the key to precision agriculture is crop productivity and its sustainability, the decision tree classification algorithms could support in predicting the circumstances responsible for low or medium or high sugarcane crop yield under certain climatic parameters.

Acknowledgements

The first author would like to thank the management of Kalasalingam Academy of Research and Education for providing fellowship. Authors also grant gratitude towards the Director, ICAR-Sugarcane Breeding Institute, Coimbatore for providing the necessary data to carry out the research work.

References

Aksu, G. and N. Dogan (2019). Comparison of decision trees used in data mining. *Journal of Education and Instruction.,* **9(4):** 1183.

Bikku, T., S.N. Rao and R.A. Akepogu (2016). Hadoop based Feature Selection and Decision Making Models on Big Data", *International Journal of Science and Technology.*, **9(10):** 1-6.

Dai, W. and W. Ji (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm", *International Journal of Database Theory and Application.*, **7(11):** 50-60.

Glory, A.H., R. Nithya and I.S. Jeyapaul (2015). Comparing C4.5 and MST Classifier Using MapReduce, *International Research Journal of Engineering and Technology.*, **2(2):** 1-4.

Krishna Kumar, V.S. and P. Kiruthika (2015). An Overview of Classification Algorithm in Data Mining, *International Journal of Advanced Research in Computer and Communication Engineering.*, **4(12):** 255-257.

Rale, N., R. Solanki, D. Bein J. andro-Vasko and W. Bein (2019). Prediction of Crop Cultivation. In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 0227-0232. IEEE.

Revathi, P., R. Revathi and M. Hemalatha (2011). Comparative Study of Knowledge in Crop Diseases Using Machine Learning Techniques, *International Journal of Computer Science and Information Technologies.*, **2(5):** 2180-2182.

Revathy, R. and R. Lawrance (2017). Classifying Crop Pest Data using C4. 5 algorithm' in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 1-6.

Revathy, R. and R. Lawrance (2017). Comparative Analysis of C4. 5 and C5. 0 Algorithms on Crop Pest Data, *International Journal of Innovative Research in Computer and Communication Engineering.*, **5(1):** 50-58.

Revathy, R., S. Balamurali and R. Lawrance (2019). Classifying Agricultural Crop Pest Data Using Hadoop MapReduce Based C5. 0 Algorithm. *Journal of Cyber Security and Mobility.*, **8(3):** 393-408.

Rosario, F.S. and K. Thangadurai (2015). RELIEF: Feature Selection Approach", *International Journal of Innovative Research & Development.*, **4(11)**.

Sahu, S., M. Chawla and N. Khare (2019). Viable Crop Prediction Scenario in BigData Using a Novel Approach. In Emerging Technologies in Data Mining and Information Security, 165-177. Springer, Singapore.

Sutha, S. and J.J. Tamilselvi (2015). A Review of Feature Selection Algorithms for Data Mining Techniques", *International Journal on Computer Science and Engineering.*, **7(6):** 62-67.

Veenadhari, S., B. Mishra and C.D. Singh (2011). Soybean Productivity Modelling using Decision Tree Algorithms. *International Journal of Computer Applications.*, **27(7):** 11-15.

Yang, T. and H.A. HiongNgu (2016). Implementation of Decision Tree Using Hadoop Map Reduce, *International Journal of Biomedical Data Mining.*, **6(1):** 1-4.